

A Semi-automated Pipeline for Phenotyping Diabetes Complications in UK Biobank : Generating and Analyzing Interval-censored Time-to-event Data

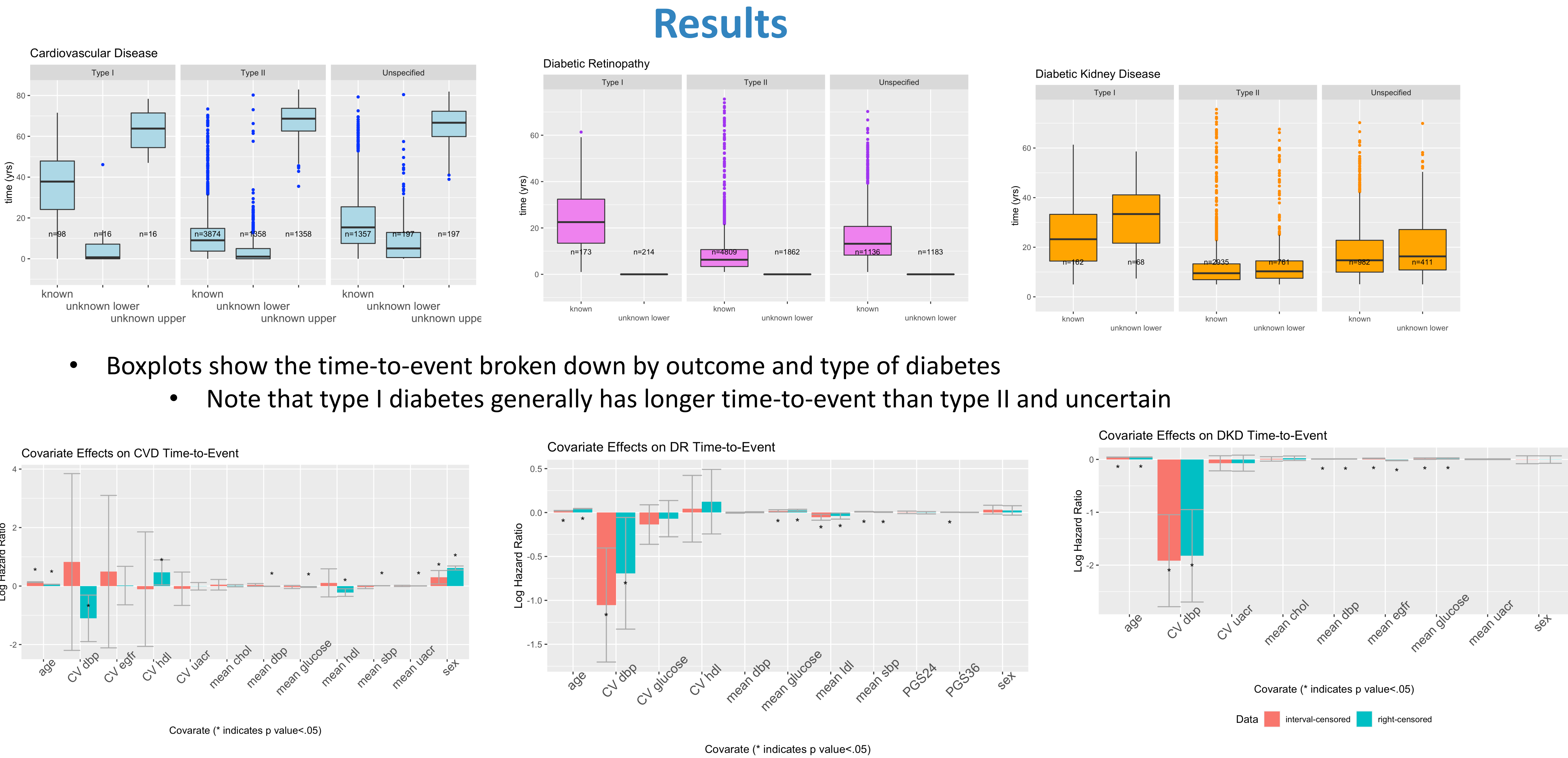
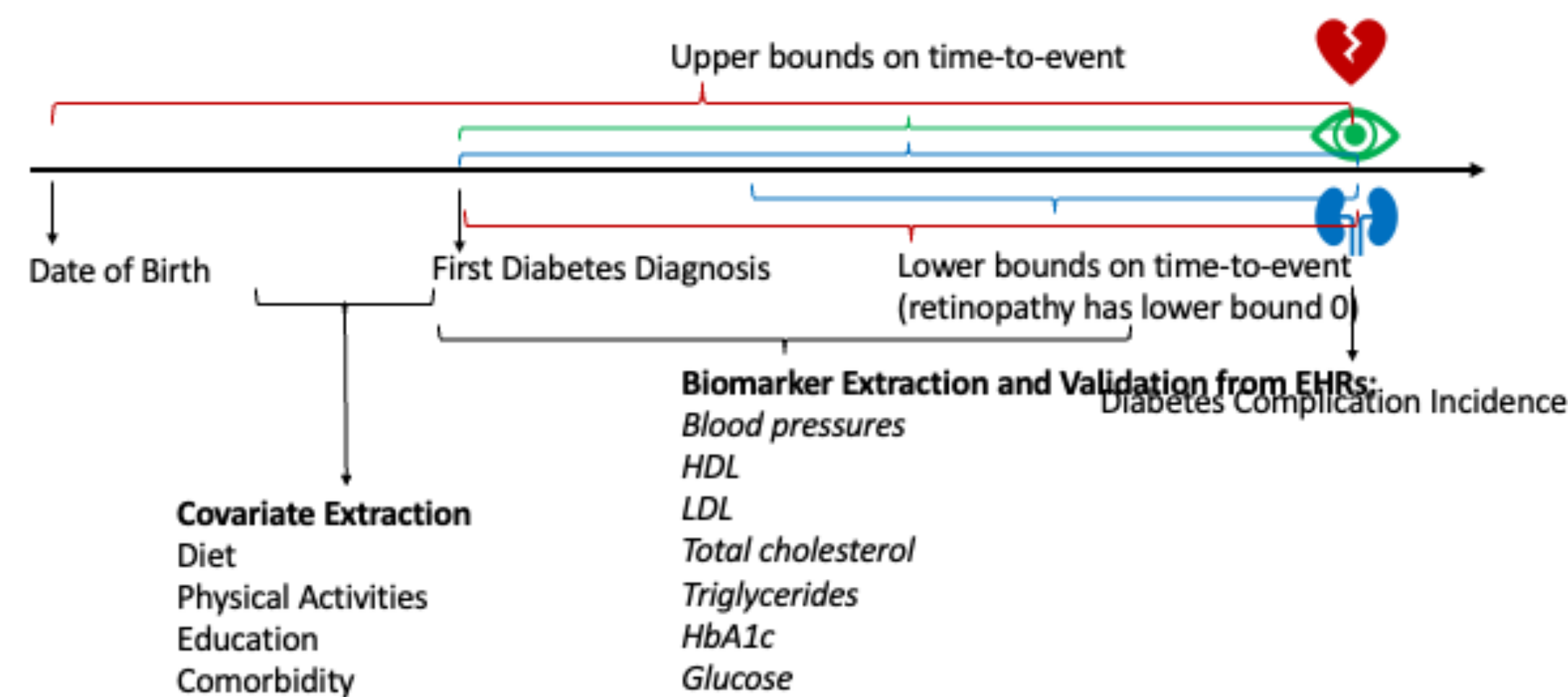
KELLY JONES, Do Hyun Kim, Aubrey Jensen, Hua Zhou, Jin Zhou
Bruins in Genomics Summer Program, OpenMendel Lab

Background

- Electronic health records (EHRs) from the UK Biobank study allow the observation of development of diabetes complications over time
- Aims to create a pipeline to generate time-to-event data from raw data on three diabetes complications

Methods

- Use EHRs codes and dates to phenotype outcomes and compute time-to-event
- Use biomarker and demographic data to find covariates
- Interval-censored data has upper and lower bounds on time-to-event
- Analyze interval and right-censored data with covariates to predict risk with icenReg, miceRanger, and survival R packages



- Boxplots show the time-to-event broken down by outcome and type of diabetes
 - Note that type I diabetes generally has longer time-to-event than type II and uncertain
- Bar graphs show the effect sizes of selected covariates on time-to-event broken down by type of data and outcome
 - Age has a significant effect for all outcomes with both datasets
 - More agreement on which covariates are significant between interval and right-censored data for DKD and DR

Abbrevia tion	Meaning
CVD	Cardiovascular Disease
DR	Diabetic Retinopathy
DKD	Diabetic Kidney Disease
CV	Coefficient of Variation
sbp	Systolic blood pressure
dpb	Diastolic blood pressure
egfr	Estimated glomerular filtration rate
uacr	Urine albumin creatinine ratio
hdl/dl	High/low-density lipoprotein cholesterol

Conclusion

- The semi-automated pipeline can provide valuable, time-to-event data and covariates to study diabetes complciations
- The pipeline can be modified to produce both right-censored and interval-censored data based on HER information
- Interval-censored data shows many of the same significant covariates as right-censored data for DR and DKD but the datasets differ for CVD

References

- Darke, et al. “Curating a longitudinal research resource using linked primary care EHR data—a UK Biobank case study”. 2022.
- Bansal & Heagerty. “A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes”. 2019